Anlage 1: Modulkonzept zu k-means Clustering

Förderkennzeichen:	01IS17073
Vorhabenbezeichnung:	Verbundprojekt: COSY-Entwicklung von sieben Praxis-Versuchen zum Thema Datenanalyse und Maschinelles Lernen an zwei Hochschulstandorten
Zugehörigkeit:	Modulkonzept zu Versuch 1 - Schallortung

Allgemeines:

Das k-Means Clustering ist ein vergleichsweise einfaches Verfahren zur Gruppenbildung aus nicht gelabelten Daten, weshalb es sich besonders zur Einführung in das Machine Learning eignet. Es wird häufig verwendet, um aus einem übergebenen Datensatz k Gruppen nach einem zuvor definierten Kriterium zu bilden.
Im k-Means Algorithmus wird eine übergebene Menge $\mathcal{M} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ von n Datenvektoren $\mathbf{x}_i \in \mathbb{R}^d$ in k Gruppen $\mathcal{C}_1, \ldots, \mathcal{C}_k$ geteilt. Dazu werden zunächst k Datenvektoren als Referenzvektoren $\mathbf{r}^{neu}$ aus $\mathcal{M}$ gewählt. Diese können im Rahmen der Toolbox zufällig, oder auf verschiedene andere Arten initialisiert werden [2]. Der Algorithmus ordnet die Datenvektoren $\mathbf{x}_1, \ldots, \mathbf{x}_n$ nach einem Kriterium, z. B. der geringsten euklidischen Distanz, den gewählten Referenzvektoren zu. Von den so entstandenen Clustern wird je ein neuer repräsentativer Referenzvektor $\mathbf{r}_j^{neu} \in \mathbb{R}^d$ mit $j = 1, \ldots, k$ durch

$$\mathbf{r}_j^{neu} = \text{argmin}_{\mathbf{r}_j \in \mathbb{R}^d} \sum_{\mathbf{x}_i \in \mathcal{C}_j} d(\mathbf{x}_i, \mathbf{r}_j) \tag{0.1}$$

bestimmt, der in der Summe den geringsten Abstand zu allen Datenvektoren seines Clusters hat. Vor der Bestimmung des neuen Referenzvektors werden die ermittelten Prototypen $\mathbf{r}_j^{neu}$ in $\mathbf{r}_j^{alt}$ gespeichert, um die optimierten Referenzvektoren $\mathbf{r}_j^{neu}$ mit den Referenzvektoren vor der Optimierung $\mathbf{r}_j^{alt}$ vergleichen zu können. Sind die neuen Referenzvektoren weniger als eine definierte Schwelle $d_{Schwellwert}$ von den bestehenden Referenzvektoren entfernt (hier $d_{Schwellwert} = 0, \mathbf{r}_j^{neu} \neq \mathbf{r}_j^{alt}$), wird das Verfahren beendet und der k-Means-Algorithmus liefert die ermittelten Referenzvektoren zurück. Wurde der Schwellwert noch nicht erreicht werden die Datenvektoren $\mathbf{x}_1, \ldots, \mathbf{x}_n$ wiederum den neuen Referenzvektoren zugeordnet und der Referenzvektor im Anschluss optimiert. Dieses Verfahren wird solange wiederholt bis der Schwellwert unterschritten wird.[1]
Die Optimierung erfolgt im Falle der euklidischen Distanz durch Bildung des Mittelwerts (mean) der Datenvektoren eines Clusters, was der Division der Summierung der zum Cluster gehörenden Vektoren durch deren Anzahl entspricht. Genauer gilt im Falle der euklidischen Distanz:

$$\text{argmin}_{\mathbf{r}_j \in \mathbb{R}^d} \sum_{\mathbf{x}_i \in \mathcal{C}_j} d(\mathbf{x}_i, \mathbf{r}_j) = \text{argmin}_{\mathbf{r}_j \in \mathbb{R}^d} \sum_{\mathbf{x}_i \in \mathcal{C}_j} \|\mathbf{x}_i - \mathbf{r}_j\|^2$$

Die Minimierung der Summe der Abstände erfolgt durch Nullsetzen der Ableitung

$$\frac{\partial \sum_{\mathbf{x}_i \in \mathcal{C}_j} \|\mathbf{x}_i - \mathbf{r}_j\|^2}{\partial d} = 0 \iff 2 \sum_{\mathbf{x}_i \in \mathcal{C}_j} \mathbf{x}_i - \mathbf{r}_j = 0. \tag{0.2}$$

Da $r_j = \frac{\sum_{\mathbf{x}_i \in \mathcal{C}_j} \mathbf{x}_i}{\sum_{\mathbf{x}_i \in \mathcal{C}_j} 1} = \frac{\sum_{\mathbf{x}_i \in \mathcal{C}_j} \mathbf{x}_i}{|\mathcal{C}_j|}$ die Gleichung (0.2) erfüllt, wird der folgende Algorithmus k-Means genannt.

k-Means Algorithmus:

Algorithmus 1 stellt den für den Versuch verwendeten Pseudocode zum k-Means Clustering dar.[1]

Algorithm 1 k-Means: $k\text{Means}(\mathcal{M})$
Input: Menge von n Datensätzen $\mathcal{M} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$
Output: $k < n$ Prototypen $(\mathbf{r}_1, \ldots, \mathbf{r}_k)$ für die Datensätze

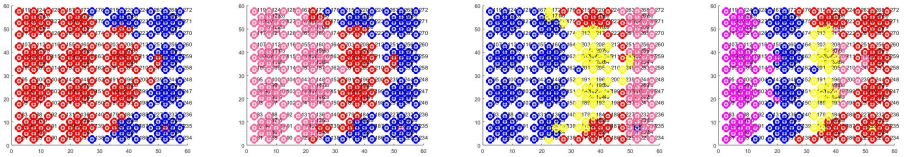
```
for j = 1, ..., k do
    r_j^neu = x_j % Zufällige Auswahl
end for
r^alt = [∞, ..., ∞]^T
r^neu = [r_1^neu, ..., r_k^neu]^T
while r^neu ≠ r^alt do
    for i = 1, ..., n do
        best = ∞
        for j = 1, ..., k do
            dist = d(x_i, r_j^neu)
            if dist < best then
                best = dist
                representative(x_i) = j
            end if
        end for
        add x_i to C_representative(x_i)
    end for
    r^alt = r^neu
    for j = 1, ..., k do
        r_j^neu = argmin_{r_j ∈ R^d} Σ_{x_i ∈ C_j} d(x_i, r_j)
    end for
end while
return ((r_1^neu, ..., r_k^neu))
```

Versuch 1: Schallortung liefert gute Daten für die Verwendung von k-Means. Aus diesem Grund bildet Versuch 1 den Rahmen für das Modulkonzept k-Means.

Aufgaben:

Die Studenten rufen den k-Means Algorithmus auf ungelabelte Raumimpulsantworten, die über den gesamten Tisch aufgenommen wurden, mit verschiedenen Konfigurationen auf, um die unterschiedlichen Ergebnisse zu beobachten und zu interpretieren. Dabei sollte ein fehlerhafter Datensatz auffallen, der nach Begutachtung des "Plots" korrigierbar ist. Die wesentlichen Aufgaben sind:

1. Aufruf der k-Means Toolbox mit unterschiedlicher Clusteranzahl k.



2. Visualisierung und Interpretation der Ergebnisse

3. Erkennen eines fehlerhaften Datensatzes anhand der Ergebnisse.

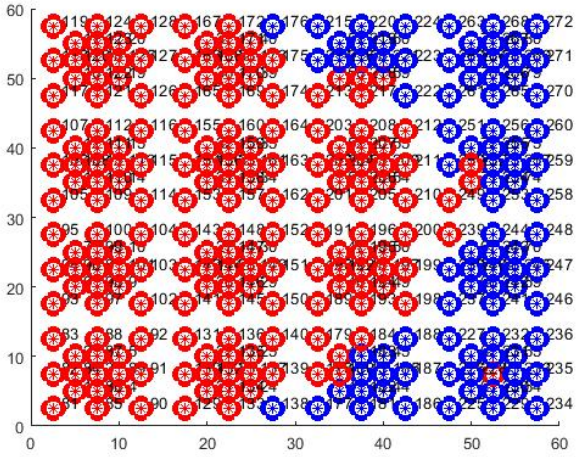
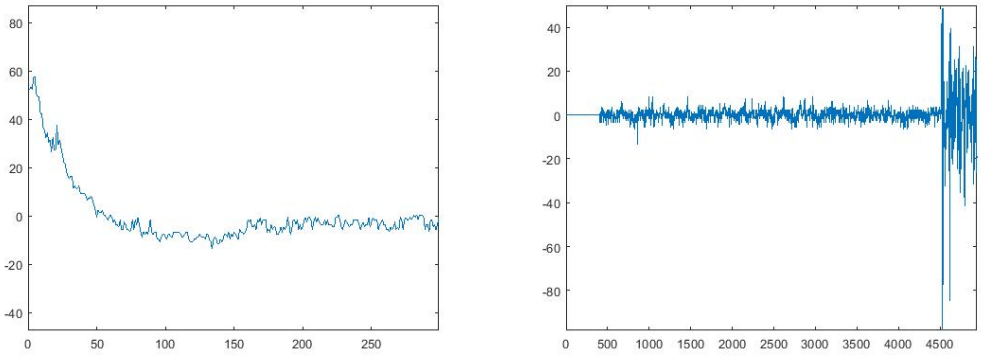


Abbildung 1: Datensatz 63, der rote Punkt in der rechten unteren Ecke sollte auffallen.

4. Visualisierung des fehlerhaften Datensatzes und Korrektur



Im linken Bild ist erkennbar, dass die Schwelle von |50| bereits direkt zu Beginn überschritten

wird. Dies führt zum Filtern der ersten 2000 Messwerte, die nur Rauschen enthalten (siehe rechtes Bild).

5. Herleitung der Formel für die Berechnung von $\mathbf{r}_j^{new} = \text{argmin}_{\mathbf{r}_j \in \mathbb{R}^d} \sum_{\mathbf{x}_i \in \mathcal{C}_j} d(\mathbf{x}_i, \mathbf{r}_j)$ mit euklidischer Distanz.

6. Eigenständige Implementierung des Algorithmus mit Euklidischer Distanz.

7. Vergleich der Ergebnisse mit und ohne Toolbox.

Ziele:

- Die Studenten kennen optionale Konfigurationen von k-Means in der Toolbox und sind in der Lage Datensätze zu Analysezwecken zu visualisieren.
- Die Studenten erkennen das Ziel von unsupervised learning und stellen fest, dass Schallsignale als Trainingsdaten für eine Lokalisierung verwendet werden können, da der Algorithmus den Tisch auf Basis der Schallsignale in zusammenhängende Bereiche unterteilt.
- Die Studenten können aussagekräftige Cluster-Bildungen von nicht-aussagekräftigen unterscheiden und hinterfragen die Ergebnisse des Algorithmus.
- Die Studenten verstehen die Funktionsweise des Algorithmus und sind in der Lage ihn eigenständig zu reproduzieren.

Besonderheit:

Anstatt nur die Identifikation von Gruppen durch den k-Means Algorithmus zu verdeutlichen, wird durch die Gruppierung die praktische Frage: "Können Raumimpulsantworten für die Lokalisierung genutzt werden?" beantwortet. Dadurch ist eine tiefere Interpretation der Ergebnisse möglich. Zudem wird durch das Finden des fehlerhaften Datensatzes ein Erfolgserlebnis vermittelt.

Literatur

[1] Matthias Dziubany, Rüdiger Machhamer, Hendrik Laux, Anke Schmeink, Klaus-Uwe Gollmer, Guido Burger, and Guido Dartmann. Machine learning based indoor localization using a representative k-nearest-neighbor classifier on a low-cost iot-hardware.

[2] MathWorks. k-means clustering - matlab kmeans - mathworks deutschland, 2018.